Combining EfficientNet and Vision Transformers for Video Deepfake Detection

Year: 2021 | Citations: 210 | Authors: D. Coccomini, Nicola Messina, C. Gennaro, F. Falchi

---

Abstract

Deepfakes are the result of digital manipulation to forge realistic yet fake imagery. With the astonishing advances in deep generative models, fake images or videos are nowadays obtained using variational autoencoders (VAEs) or Generative Adversarial Networks (GANs). These technologies are becoming more accessible and accurate, resulting in fake videos that are very difficult to be detected. Traditionally, Convolutional Neural Networks (CNNs) have been used to perform video deepfake detection, with the best results obtained using methods based on EfficientNet B7. In this study, we focus on video deep fake detection on faces, given that most methods are becoming extremely accurate in the generation of realistic human faces. Specifically, we combine various types of Vision Transformers with a convolutional EfficientNet B0 used as a feature extractor, obtaining comparable results with some very recent methods that use Vision Transformers. Differently from the state-of-the-art approaches, we use neither distillation nor ensemble methods. Furthermore, we present a straightforward inference procedure based on a simple voting scheme for handling multiple faces in the same video shot. The best model achieved an AUC of 0.951 and an F1 score of 88.0%, very close to the state-of-the-art on the DeepFake Detection Challenge (DFDC).